PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-162986
(43)Date of publication of application : 07.06.2002

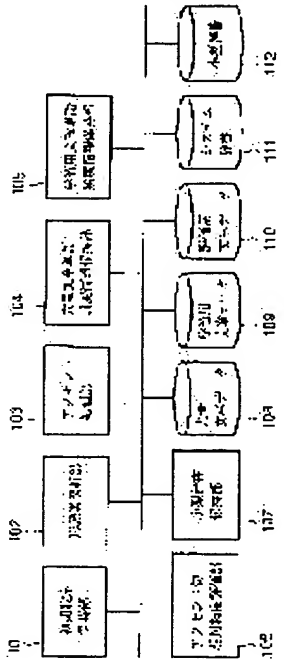
(51)Int.Cl. G10L 13/08
G10L 13/00
G10L 15/06
G10L 13/04

(21)Application number : 2000-360207
(22)Date of filing : 27.11.2000

(71)Applicant : CANON INC
(72)Inventor : AIZAWA MICHIO

(54) DEVICE AND METHOD FOR INFORMATION PROCESSING AND COMPUTER-READABLE MEMORY

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a device and a method for information processing and a computer-readable memory which can generate a highly flexible and precise dictionary for voice synthesis.
SOLUTION: An initial setting hold part 101 specifies the number of words constituting a small-size dictionary 112. A small-size dictionary hold part 107 constitutes the small-size dictionary 112 according to learning document data 109 for determining the words constituting the small-size dictionary 112 and large-amount document data 108. An accent processing part 103 performs accent processing for evaluation document data 110 for evaluating the small-size dictionary 112 by using a system dictionary 111 to output 1st accent phrase information and performs accent processing by using the small-size dictionary to output 2nd accent phrase information. An accent phrase relative precision evaluation part 106 calculates the relative precision of the 2nd accent phrase information to the 1st accent phrase information. Then the small-size dictionary hold part 107 performs management while making the relative precision and small-size dictionary 112 correspond to each other.



LEGAL STATUS

[Date of request for examination] 10.06.2004
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－162986
（Ｐ２００２－１６２９８６Ａ）
(43)公開日　平成14年６月７日(2002.6.7)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
G10L 13/08		G10L 3/00	H	5D015
13/00			Q	5D045
15/06			521 C	
13/04			521 F	
		5/02	G	

審査請求　未請求　請求項の数11　ＯＬ　（全８頁）

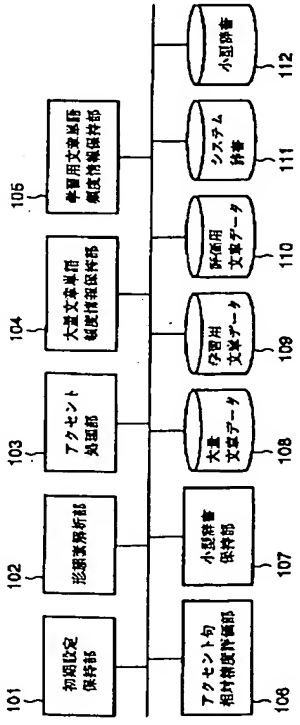
(21)出願番号　特願2000－360207(Ｐ2000－360207)
(22)出願日　平成12年11月27日(2000.11.27)

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号
(72)発明者　相澤　道雄
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内
(74)代理人　100076428
弁理士　大塚　康徳　（外２名）

Ｆターム(参考)　5D015 GG01
5D045 AA09 AB01

(54)【発明の名称】情報処理装置及びその方法、コンピュータ可読メモリ

(57)【要約】
【課題】　汎用性が高く、かつ精度の良い音声合成用辞書を作成することができる情報処理装置及びその方法、コンピュータ可読メモリを提供する。
【解決手段】　初期設定保持部１０１で、小型辞書１１２を構成する単語数を指定する。次に、小型辞書保持部１０７で、小型辞書１１２を構成する単語を決定するための学習用文書データ１０９と、大量文章データ１０８に基づいて、該小型辞書１１２を構成する。アクセント処理部１０３は、小型辞書１１２を評価するための評価用文章データ１１０に対し、システム辞書１１１を用いてアクセント処理を行ない、第１アクセント句情報を出力し、また、評価用文章データ１１０に対し、小型辞書１１２を用いてアクセント処理を行い、第２アクセント句情報を出力する。そして、アクセント句相対精度評価部１０６は、第２アクセント句情報の第１アクセント句情報に対する相対精度を算出する。そして、小型辞書保持部１０７は、相対精度と小型辞書１１２を対応づけて管理する。



1

【特許請求の範囲】

【請求項１】　予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理装置であって、

前記小型辞書を構成する単語数を指定する指定手段と、

前記小型辞書を構成する単語を決定するための学習用文書データと、大量文章データに基づいて、該小型辞書を構成する構成手段と、

前記小型辞書を評価するための評価用文章データに対し、前記システム辞書を用いてアクセント処理を行ない、第１アクセント句情報を出力する第１アクセント処理手段と、

前記評価用文章データに対し、前記小型辞書を用いてアクセント処理を行い、第２アクセント句情報を出力する第２アクセント処理手段と、

前記第２アクセント句情報の前記第１アクセント句情報に対する相対精度を算出する算出手段と、

前記相対精度と前記小型辞書を対応づけて管理する管理手段とを備えることを特徴とする情報処理装置。

【請求項２】　前記指定手段は、前記小型辞書を構成する全単語数と、該全単語数の内、前記学習用文書データから得られる単語から選ぶ単語数との組を指定することを特徴とする請求項１に記載の情報処理装置。

【請求項３】　前記構成手段は、前記学習用文書データ及び前記大量文書データそれぞれに対し形態素解析を行う形態素解析手段とを備え、

前記形態素解析手段の形態素解析結果に基づいて、前記小型辞書を構成することを特徴とする請求項１に記載の情報処理装置。

【請求項４】　前記相対精度は、前記第２アクセント情報が前記第１アクセント情報と一致する度合を示すことを特徴とする請求項１に記載の情報処理装置。

【請求項５】　前記学習用文章データ及び前記評価用文章データは、分野毎に存在することを特徴とする請求項１に記載の情報処理装置。

【請求項６】　予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理方法であって、

前記小型辞書を構成する単語数を指定する指定工程と、

前記小型辞書を構成する単語を決定するための学習用文書データと、大量文章データに基づいて、該小型辞書を構成する構成工程と、

前記小型辞書を評価するための評価用文章データに対し、前記システム辞書を用いてアクセント処理を行ない、第１アクセント句情報を出力する第１アクセント処理工程と、

前記評価用文章データに対し、前記小型辞書を用いてアクセント処理を行い、第２アクセント句情報を出力する第２アクセント処理工程と、

前記第２アクセント句情報の前記第１アクセント句情報に対する相対精度を算出する算出工程と、

前記相対精度と前記小型辞書を対応づけて記憶媒体に管理する管理工程とを備えることを特徴とする情報処理方法。

2

【請求項７】　前記指定工程は、前記小型辞書を構成する全単語数と、該全単語数の内、前記学習用文書データから得られる単語から選ぶ単語数との組を指定することを特徴とする請求項６に記載の情報処理方法。

【請求項８】　前記構成工程は、前記学習用文書データ及び前記大量文書データそれぞれに対し形態素解析を行う形態素解析工程とを備え、

前記形態素解析工程の形態素解析結果に基づいて、前記小型辞書を構成することを特徴とする請求項６に記載の情報処理方法。

【請求項９】　前記相対精度は、前記第２アクセント句情報が前記第１アクセント情報と一致する度合を示すことを特徴とする請求項６に記載の情報処理方法。

【請求項１０】　前記学習用文章データ及び前記評価用文章データは、分野毎に存在することを特徴とする請求項６に記載の情報処理方法。

【請求項１１】　予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理のプログラムコードが格納されたコンピュータ可読メモリであって、

前記小型辞書を構成する単語数を指定する指定工程のプログラムコードと、

前記小型辞書を構成する単語を決定するための学習用文書データと、大量文章データに基づいて、該小型辞書を構成する構成工程のプログラムコードと、

前記小型辞書を評価するための評価用文章データに対し、前記システム辞書を用いてアクセント処理を行ない、第１アクセント句情報を出力する第１アクセント処理工程のプログラムコードと、

前記評価用文章データに対し、前記小型辞書を用いてアクセント処理を行い、第２アクセント句情報を出力する第２アクセント処理工程のプログラムコードと、

前記第２アクセント句情報の前記第１アクセント句情報に対する相対精度を算出する算出工程のプログラムコードと、

前記相対精度と前記小型辞書を対応づけて記憶媒体に管理する管理工程とを備えることを特徴とするコンピュータ可読メモリ。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理装置及びその方法、コンピュータ可読メモリに関するものである。

【０００２】

【従来の技術】音声合成システムや音声認識システム等

では、音声認識用としてシステムの持つ数十万語レベルのシステム辞書から数千語～数万語の単語を取り出したサブセットの小型辞書がよく使われる。これは、実行速度の向上やメモリサイズの削減を目的としている。

【０００３】例えば、コンピュータ関係の文書と経済関係の文章では、異なる単語が出現する。しかし、小型辞書では、語数を制限しているため様々な分野の単語（特に、専門用語等）を同時に収録することができない。そのため、コンピュータ関係、経済関係等の適用分野に合わせて小型辞書に登録する選択単語を変化させる必要がある。

【０００４】適用分野に応じて、小型辞書に登録する単語を選択する最も簡単な方法は、その適用分野の大量の文章から単語の出現頻度を計算し、出現頻度の上位から設定した語数までの単語を選ぶ方法がある。しかし、一般に適用分野に応じた文章を大量に集めることは困難である。そこで、新聞記事等の大量に入手できる文章と適用分野の少量の文章に基づいて、小型辞書に登録する単語を選択する方法を行っている。

【０００５】例えば、特開２０００－７５８９２号では、大量にある過去のニュース原稿と、少量の最近のニュース原稿から、小型辞書に登録する単語を選択している。

【０００６】

【発明が解決しようとする課題】しかしながら、従来の特開２０００－７５８９２号で開示される方法は、単語選択の評価基準としてパープレキシティを用いている。パープレキシティはｎグラムモデルを用いた統計的手法に関係する値である。そのため、統計的手法を用いていないシステムにおいては有効性が低い。

【０００７】本発明は上記の課題を解決するためになされたものであり、汎用性が高く、かつ精度の良い音声合成用辞書を作成することができる情報処理装置及びその方法、コンピュータ可読メモリを提供することを目的とする。

〔発明の詳細な説明〕

【発明の属する技術分野】本発明は、予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理装置及びその方法、コンピュータ可読メモリに関するものである。

【０００８】

【従来の技術】音声合成システムや音声認識システム等では、音声認識用としてシステムの持つ数十万語レベルのシステム辞書から数千語～数万語の単語を取り出したサブセットの小型辞書がよく使われる。これは、実行速度の向上やメモリサイズの削減を目的としている。

【０００９】例えば、コンピュータ関係の文書と経済関係の文章では、異なる単語が出現する。しかし、小型辞書では、語数を制限しているため様々な分野の単語（特に、専門用語等）を同時に収録することができない。そのため、コンピュータ関係、経済関係等の適用分野に合わせて小型辞書に登録する選択単語を変化させる必要がある。

【００１０】適用分野に応じて、小型辞書に登録する単語を選択する最も簡単な方法は、その適用分野の大量の文章から単語の出現頻度を計算し、出現頻度の上位から設定した語数までの単語を選ぶ方法がある。しかし、一般に適用分野に応じた文章を大量に集めることは困難である。そこで、新聞記事等の大量に入手できる文章と適用分野の少量の文章に基づいて、小型辞書に登録する単語を選択する方法を行っている。

【００１１】例えば、特開２０００－７５８９２号では、大量にある過去のニュース原稿と、少量の最近のニュース原稿から、小型辞書に登録する単語を選択している。

【００１２】

【発明が解決しようとする課題】しかしながら、従来の特開２０００－７５８９２号で開示される方法は、単語選択の評価基準としてパープレキシティを用いている。パープレキシティはｎグラムモデルを用いた統計的手法に関係する値である。そのため、統計的手法を用いていないシステムにおいては有効性が低い。

【００１３】本発明は上記の課題を解決するためになされたものであり、汎用性が高く、かつ精度の良い音声合成用辞書を作成することができる情報処理装置及びその方法、コンピュータ可読メモリを提供することを目的とする。

【００１４】

【課題を解決するための手段】上記の目的を達成するための本発明による情報処理装置は以下の構成を備える。即ち、予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理装置であって、前記小型辞書を構成する単語数を指定する指定手段と、前記小型辞書を構成する単語を決定するための学習用文書データと、大量文章データに基づいて、該小型辞書を構成する構成手段と、前記小型辞書を評価するための評価用文章データに対し、前記システム辞書を用いてアクセント処理を行ない、第１アクセント句情報を出力する第１アクセント処理手段と、前記評価用文章データに対し、前記小型辞書を用いてアクセント処理を行い、第２アクセント句情報を出力する第２アクセント処理手段と、前記第２アクセント句情報の前記第１アクセント句情報に対する相対精度を算出する算出手段と、前記相対精度と前記小型辞書を対応づけて管理する管理手段とを備える。

【００１５】また、好ましくは、前記指定手段は、前記小型辞書を構成する全単語数と、該全単語数の内、前記学習用文書データから得られる単語から選ぶ単語数との組を指定する。

【００１６】また、好ましくは、前記構成手段は、前記学習用文書データ及び前記大量文書データそれぞれに対

し形態素解析を行う形態素解析手段とを備え、前記形態素解析手段の形態素解析結果に基づいて、前記小型辞書を構成する。

【００１７】また、好ましくは、前記相対精度は、前記第２アクセント句情報が前記第１アクセント情報と一致する度合を示す。

【００１８】また、好ましくは、前記学習用文章データ及び前記評価用文章データは、分野毎に存在する。

【００１９】上記の目的を達成するための本発明による情報処理方法は以下の構成を備える。即ち、予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理方法であって、前記小型辞書を構成する単語数を指定する指定工程と、前記小型辞書を構成する単語を決定するための学習用文書データと、大量文章データに基づいて、該小型辞書を構成する構成工程と、前記小型辞書を評価するための評価用文章データに対し、前記システム辞書を用いてアクセント処理を行ない、第１アクセント句情報を出力する第１アクセント処理工程と、前記評価用文章データに対し、前記小型辞書を用いてアクセント処理を行い、第２アクセント句情報を出力する第２アクセント処理工程と、前記第２アクセント句情報の前記第１アクセント句情報に対する相対精度を算出する算出工程と、前記相対精度と前記小型辞書を対応づけて記憶媒体に管理する管理工程とを備える。

【００２０】上記の目的を達成するための本発明によるコンピュータ可読メモリは以下の構成を備える。即ち、予め登録されているシステム辞書に基づいて分野毎の小型辞書を作成する情報処理のプログラムコードが格納されたコンピュータ可読メモリであって、前記小型辞書を構成する単語数を指定する指定工程のプログラムコードと、前記小型辞書を構成する単語を決定するための学習用文書データと、大量文章データに基づいて、該小型辞書を構成する構成工程のプログラムコードと、前記小型辞書を評価するための評価用文章データに対し、前記システム辞書を用いてアクセント処理を行ない、第１アクセント句情報を出力する第１アクセント処理工程のプログラムコードと、前記評価用文章データに対し、前記小型辞書を用いてアクセント処理を行い、第２アクセント句情報を出力する第２アクセント処理工程のプログラムコードと、前記第２アクセント句情報の前記第１アクセント句情報に対する相対精度を算出する算出工程のプログラムコードと、前記相対精度と前記小型辞書を対応づけて記憶媒体に管理する管理工程とを備える。

【００２１】

【発明の実施の形態】以下、図面を参照して本発明の好適な実施形態を詳細に説明する。

【００２２】図１は本実施形態の情報処理装置の構成を示すブロック図である。

【００２３】図１において、ＣＰＵ１１０１はメインバス１１０６を介して情報処理装置１０００全体の制御を

実行するとともに、情報処理装置１０００の外部に接続される入力装置１１１０（例えば、マイク、イメージスキャナ、記憶装置、ネットワーク回線を介して接続される他の情報処理装置、電話回線を介して接続されるファクシミリ等）を入力Ｉ／Ｆ（インタフェース）１１０４を介して制御する。また、情報処理装置１０００の外部に接続される出力装置１１１１（例えば、スピーカ、プリンタ、モニタ、ネットワーク回線を介して接続される他の情報処理装置、電話回線を介して接続されるファクシミリ等）を出力Ｉ／Ｆ１１０５を介して制御する。また、ＣＰＵ１１０１は、ＫＢＤＩ／Ｆ（キーボードインタフェース）１１０７を介して入力部（例えば、キーボード１１１２やポインティングデバイス１１１３やペン１１１４）から入力された指示に従って、音声の入力、音声認識処理、音声合成処理、等の一連の処理を実行する。更に、入力装置１１１０より入力された音声データや、キーボード１１１２やポインティングデバイス１１１３やペン１１１４を用いて作成されたデータを表示する表示部１１０９をビデオＩ／Ｆ（インタフェース）１１０８を介して制御する。

【００２４】ＲＯＭ１１０２は、ＣＰＵ１１０１の各種制御を実行する各種制御プログラムを記憶している。ＲＡＭ１１０３は、ＣＰＵ１１０１によりＯＳや本発明を実現するための制御プログラムを含むその他の制御プログラムがロードされ実行される。また、制御プログラムを実行するために用いられる各種作業領域、一時待避領域として機能する。また、入力装置１１１０より入力された画像データや、キーボード１１１２やポインティングデバイス１１１３やペン１１１４を用いて作成された画像データを、一旦、保持するＶＲＡＭ（不図示）が構成されている。

【００２５】次に、本実施形態の情報処理装置の機能構成について、図２を用いて説明する。

【００２６】図２は本実施形態の情報処理装置の機能構成を示すブロック図である。

【００２７】１０１は、本装置を動作させるための初期設定値を保持する初期設定保持部である。初期設定値は、作成対象の小型辞書１１２の登録語数とその中で学習用文書データ１１０に出現する単語の中から選択する単語数の組である。この組を複数設定することもできる。

【００２８】１０２は、入力文章データ（テキストデータ）に対し形態素解析を行う形態素解析部である。１０３は、入力文章データに対しアクセント処理を行ないアクセント句情報を出力するアクセント処理部である。アクセント句情報は、読み、アクセント句の区切り、アクセント型からなる。１０４は、大量文章データ１０８に含まれる単語の頻度情報を保持する大量文章単語頻度情報保持部である。

【００２９】１０５は、学習用文章データ１０９に含ま

れる単語の頻度情報を保持する学習用文章単語頻度情報保持部である。１０６は、アクセント句の相対精度を評価するアクセント句相対精度評価部である。ここで、アクセント句の相対精度とはシステム辞書１１１で解析したアクセント句を正解とみなした場合に、そのアクセント句と小型辞書１１２で解析したアクセント句が一致する度合（精度）を示す値のことである。

【００３０】１０７は、本装置が作成した小型辞書１１２とアクセント句相対精度の組を保持する小型辞書保持部である。１０８は、雑誌、新聞等の大量文章データである。１０９は、複数種類の適用分野の文章データに対し、各適用分野毎の小型辞書１１２の（小型辞書１１２を構成する単語を決定するための）学習用に割り当てた学習用文章データである。１１０は、複数種類の適用分野の文章データに対し、各適用分野毎の小型辞書１１２の評価用に割り当てた評価用文章データである。１１１は、本装置が持つすべての単語を含んだシステム辞書である。１１２は、後述する処理によって本装置が作成した適用分野毎の小型辞書である。

【００３１】尚、大量文章単語頻度情報保持部１０４、学習用文章単語頻度情報保持部１０５、アクセント句相対精度評価部１０６、小型辞書保持部１０７、大量文章データ１０８、学習用文章データ１０９、評価用文章データ１１０、システム辞書１１１、小型辞書１１２は、例えば、ＲＯＭ１１０２、ＲＡＭ１１０３、あるいは記憶装置として用いられる入力装置１１１０上で実現される。

【００３２】次に、本実施形態の情報処理装置で実行される処理について、図３を用いて説明する。

【００３３】図３は本実施形態の情報処理装置の処理手順を示すフローチャートである。

【００３４】まず、ステップＳ２０１で、作成対象の小型辞書１１２の登録語数とその中で学習用文章データ１０９に出現する単語の中から選ぶ単語数との組を設定する。例えば、（３００００，２００００）は、登録語数が３００００語であり、学習用文章データ１０９から選ぶ単語が２００００語の辞書を作成することを設定する。残りの１００００語は、大量文章データ１０８に出現する単語の中から選ぶ。また、この組を複数個設定してもよい。

【００３５】次に、ステップＳ２０２で、大量文章データ１０８に対しシステム辞書１１１を用いて形態素解析を行う。ステップＳ２０３で、ステップＳ２０２の形態素解析結果を用いて、大量文章データ１０８に出現する単語の頻度を計算し、頻度の高い順に大量文章単語頻度情報保持部１０４に保持する。ここで、大量文章単語頻度情報保持部１０４が保持するデータ構成例を、図４に示す。図４に示すように、大量文書単語頻度情報保持部１０４では、品詞別に、その品詞の見出しとその見出しの出現頻度を対応付けて管理している。

【００３６】次に、ステップＳ２０４で、学習用文章データ１０９に対しシステム辞書１１１を用いて形態素解析を行う。ステップＳ２０５で、ステップＳ２０４の形態素解析結果を用いて学習用文章データ１０９に出現する単語の頻度を計算し、頻度の高い順に学習用単語頻度情報保持部１０５に保持する。ここで、学習用単語頻度情報保持部１０５が保持するデータ構成例を、図５に示す。図５に示すように、学習用単語頻度情報保持部１０５では、大量文書単語頻度情報保持部１０４と同様に、品詞別に、その品詞の見出しとその見出しの出現頻度を対応付けて管理している。

【００３７】次に、ステップＳ２０６で、評価用文章データ１１０に対しシステム辞書１１１を用いてアクセント処理を行う。この処理結果は、アクセント句相対精度評価部１０６に保持する。ここで、アクセント句相対精度評価部１０６が保持するデータ構成例を、図６に示す。図６は、評価用文書データ１１０として、「半導体メモリを搭載したメモリを販売する」、「パソコンで生成したデータを再生する」、「単四型電池を使用する」に対し、システム辞書１１１を用いてアクセント処理を行った場合の処理結果を示している。処理結果は、評価用文書データの読みと、アクセント句の切れ目と、アクセント型からなり、図６では、読みを「カタカナ」、アクセント句の切れ目を「／」、アクセント型を「↓」で表している。

【００３８】次に、ステップＳ２０７で、ステップＳ２０１で設定した小型辞書１１２に登録する単語数とその中で学習用文章データ１０９に出現する単語の中から選ぶ単語数の組が残っているか否かを判定する。単語数の組が残っていない場合（ステップＳ２０７でＮＯ）、処理を終了する。一方、単語数の組が残っている場合（ステップＳ２０７でＹＥＳ）、ステップＳ２０８に進む。

【００３９】次に、ステップＳ２０８で、単語数の組を一つ取り出す。例として、単語数の組（３００００，２００００）を取り出した場合、学習用文章単語頻度情報保持部１０５に保持されている単語の中から頻度が高いもの上位２００００語を取り出し、小型辞書１１２に登録する。次に、大量文章単語頻度情報保持部１０４に保持されている単語の中で頻度の高いものから小型辞書１１２に登録する。登録対象の単語がすでに登録されている場合は、次の単語へ移る。そして、小型辞書１１２の登録語数が３００００語になるまで続ける。

【００４０】次に、ステップＳ２０８で、評価用文章データ１１０に対し小型辞書１１２を用いてアクセント処理を行う。このアクセント処理結果は、アクセント句相対精度評価部１０６に保持する。ここで、アクセント句相対精度評価部１０６が保持するデータ構成例を、図７に示す。図７は、図６と同様に、評価用文書データ１１０として、「半導体メモリを搭載したメモリを販売する」、「パソコンで生成したデータを再生する」、「単

四型電池を使用する」に対し、小型辞書１１２を用いてアクセント処理を行った場合の処理結果を示している。図７では、評価用文書データ１１０の内、「半導体メモリを搭載したメモリを販売する」のアクセント処理結果が、図６の場合と異なっていることがわかる。これは、小型辞書１１２に「半導体」という単語が登録されておらず、「半導体」を「半導」と「体」の２文字として認識されてしまった結果、その読みとして「ハンドウ」と「カラダ」が生成されていることがわかる。

【００４１】次に、ステップＳ２０９で、システム辞書１１１を用いたアクセント処理（ステップＳ２０６）の処理結果（図６）と小型辞書１１２を用いたアクセント処理（ステップＳ２０８）の処理結果（図７）を用いて、アクセント句相対精度を計算する。アクセント句相対精度の値は、次式を用いて計算する。つまり、アクセント句相対精度＝（切れ目とアクセント型が一致したアクセント句数）÷（システム辞書１１１を用いたアクセント処理の処理結果によるアクセント句数）。

【００４２】尚、切れ目とアクセント型が一致するアクセント句はＤＰマッチング等の手法を用いることで計算可能である。図８に一致するアクセント句の例を示す。図中、実線で囲まれた部分がシステム辞書１１１でアクセント処理した処理結果のアクセント句、破線で囲まれた部分が小型辞書１１２でアクセント処理した処理結果のアクセント句を示す。この図８の場合のアクセント句相対精度は、７÷１０＝０．７となる。

【００４３】次に、ステップＳ２１０で、ステップＳ２０７で選択した小型辞書１１２とステップＳ２０９で計算したアクセント句相対精度の値を小型辞書保持部１０７に保持する。

【００４４】以上の処理を繰り返すことにより、小型辞書保持部１０７に小型辞書１１２が複数個生成され、各小型辞書１１２には、それぞれのアクセント句相対精度が保持される。そして、このアクセント句相対精度により、各小型辞書１１２の性能を図ることができる。つまり、アクセント句相対精度に基づいて、ステップＳ２０１で設定する組を任意に変更することで、よりアクセント句相対精度の高い組の小型辞書１１２を構成することができる。

【００４５】以上説明したように、本実施形態によれば、新聞等の大量文章と適用分野の少量文章から、大規模なシステム辞書のサブセットとしての小型辞書を適用分野に適した形で作成することができる。また、アクセント句相対精度を評価基準としているため、辞書を利用するシステムのアルゴリズム（統計的手法、ルールベースなど）に依存しない。

【００４６】尚、本発明は、複数の機器（例えば、ホストコンピュータ、インタフェース機器、リーダ、プリンタなど）から構成されるシステムに適用しても、一つの機器からなる装置（例えば、複写機、ファクシミリ装置など）に適用してもよい。

【００４７】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読出し実行することによっても、達成されることは言うまでもない。

【００４８】この場合、記憶媒体から読出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【００４９】プログラムコードを供給するための記憶媒体としては、例えば、フロッピディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【００５０】また、コンピュータが読出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００５１】更に、記憶媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００５２】本発明を上記記憶媒体に適用する場合、その記憶媒体には、先に説明した図３に示すフローチャートに対応するプログラムコードが格納されることになる。

【００５３】

【発明の効果】以上説明したように、本発明によれば、汎用性が高く、かつ精度の良い音声合成用辞書を作成することができる情報処理装置及びその方法、コンピュータ可読メモリを提供できる。

【図面の簡単な説明】

【図１】本実施形態の情報処理装置の構成を示すブロック図である。

【図２】本実施形態の情報処理装置の機能構成を示すブロック図である。

【図３】本実施形態の情報処理装置の処理手順を示すフローチャートである。

【図４】本実施形態の大量文章単語頻度情報保持部のデータ構成例を示す図である。

11

【図５】本実施形態の学習用単語頻度情報保持部のデータ構成例を示す図である。

【図６】本実施形態の評価用文章データに対しシステム辞書を用いてアクセント処理を行った場合の処理結果を示す図である。

【図７】本実施形態の評価用文章データに対し小型辞書を用いてアクセント処理を行った場合の処理結果を示す図である。

【図８】本実施形態のアクセント句相対精度の計算方法を説明するための図である。

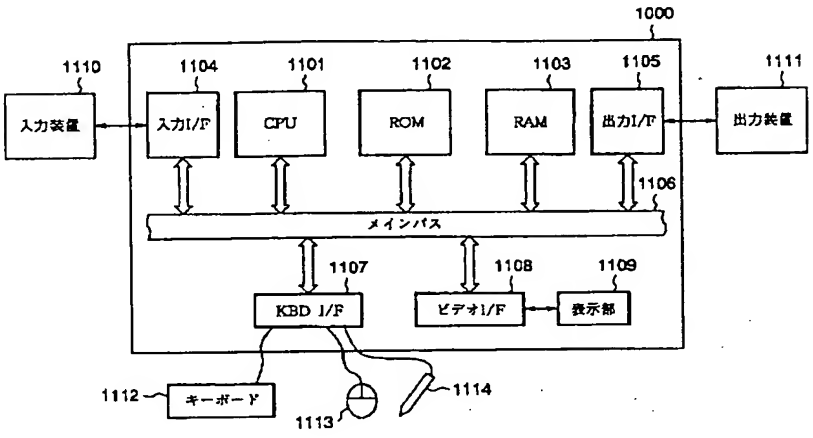
【符号の説明】

１０１　初期設定保持部

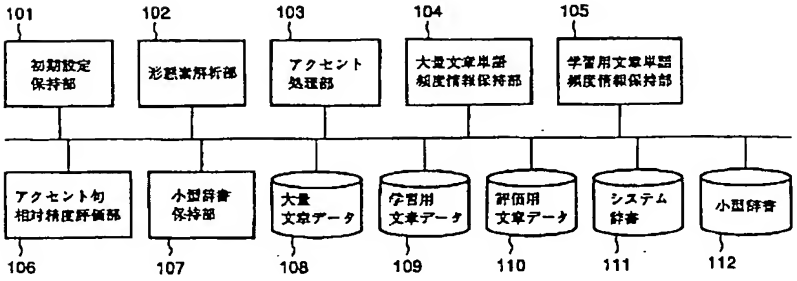
12

１０２　形態素解析部
１０３　アクセント処理部
１０４　大量文章単語頻度情報保持部
１０５　学習用文章単語頻度情報保持部
１０６　アクセント句相対精度評価部
１０７　小型辞書保持部
１０８　大量文章データ
１０９　学習用文章データ
１１０　評価用文章データ
１１１　システム辞書
１１２　小型辞書

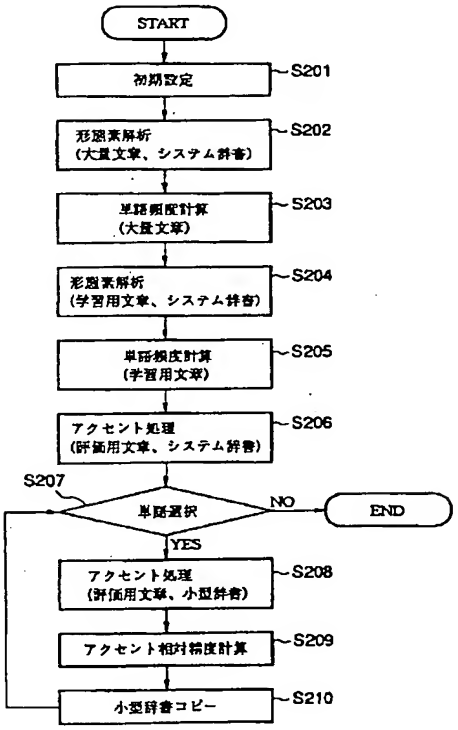
【図１】



【図２】



【図３】



【図５】

品詞	見出し	頻度
格助詞	の	22190
格助詞	を	18657
…	…	…
サ変名詞	サービス	2002
名詞	パソコン	1729
…	…	…

【図４】

品詞	見出し	頻度
格助詞	の	8389136
格助詞	を	5278827
…	…	…
固有名詞	日本	305089
助数詞	円	304169
…	…	…

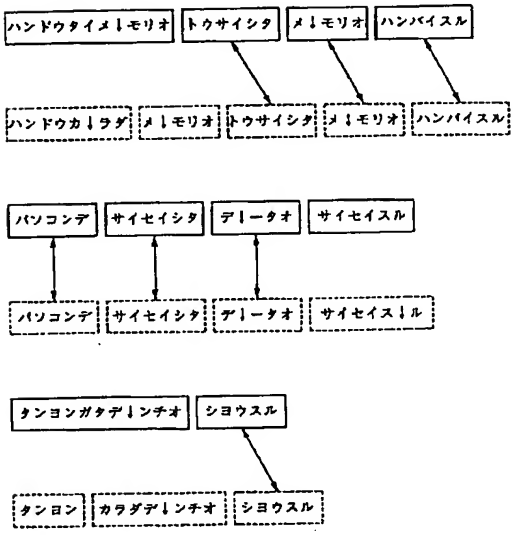
【図６】

ハンドウタイメ↓モリオ／トウサイシタ／メ↓モリヲ／ハンバイスル

パソコンデ／セイセイシタ／デ↓ータオ／サイセイスル

タンヨンガタデ↓ンチオ／シヨウスル

【図８】



【図７】

ハンドウカ↓ラダ／メ↓モリオ／トウサイシタ／メ↓モリヲ／ハンバイスル

パソコンデ／セイセイシタ／デ↓ータオ／サイセイスル

タンヨン／カタチデ↓ンチオ／シヨウスル